

UCREL Corpus Research Seminar

Lancaster University

23 May 2011

The peaks and troughs of corpus-based contextual analysis

Costas Gabrielatos, Tony McEnery, Peter Diggie & Paul Baker
(Lancaster University)

Abstract

This presentation addresses a criticism of corpus-based approaches to critical discourse studies, namely that the CL analysis does not take account of the relevant context, and shows how a preliminary corpus-based analysis can pinpoint salient contextual elements, which can inform both the CL and CDA analyses. The discussion also focuses on the importance of the statistical identification of diachronic trends (in particular, frequency peaks and troughs), and the need for high granularity in diachronic corpora. The paper aims to contribute to the synergy between CL and CDA approaches, and between qualitative and quantitative techniques in general.

The presentation uses a recently completed ESRC-funded project as a case study, The Representation of Islam in the UK Press, which used a diachronic corpus of topic-specific articles. Periods of increased frequency in the number of corpus articles were identified through a statistical analysis. These frequency peaks indicate short periods (months) of significantly increased reporting on the topic/entities in focus. These periods can then be matched with events which are expected to have triggered the increased interest. This identification has a dual benefit: a) it suggests the contextual background against which the results of the corpus analysis can be interpreted; b) it provides a reliable guide to the corpus texts that can be usefully downsampled for close (qualitative) critical discourse analysis.

Focus

- Diachronic corpus studies: relevant issues
 - Time span
 - Sampling points
 - Granularity
- Context, CDA and CL
- Identifying spikes
 - Whole corpus vs. Sub-corpora
 - Manually (impressionistically) vs. Statistically
- Utility for CDA and CL

Projects

The representation of Islam and Muslims in the UK press, 1998-2008. ESRC.

- September 2009 – August 2010.
- PI: Paul Baker; CI: Tony McEnery; RA: Costas Gabrielatos.
- Corpus: 200,000 articles; 140 million words

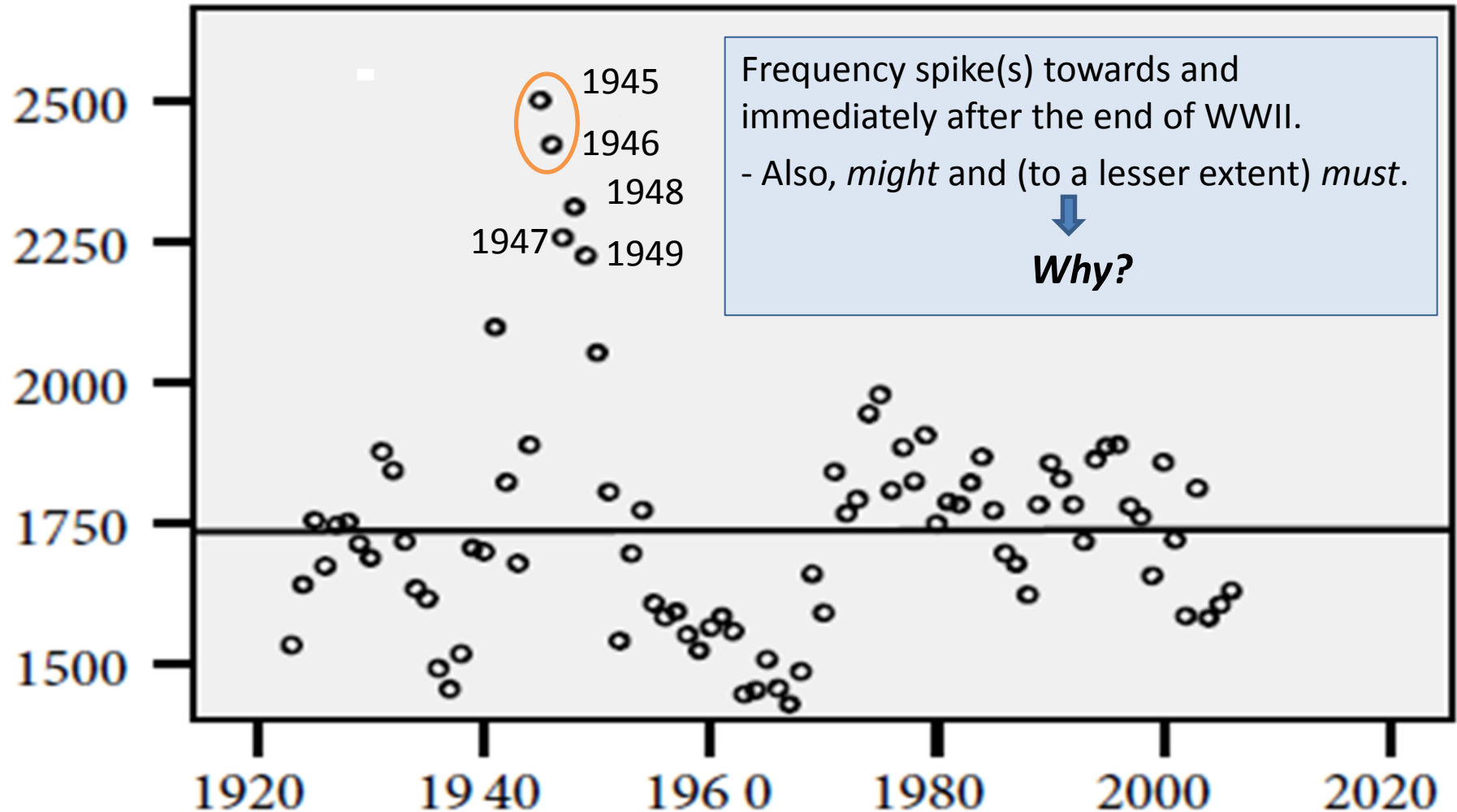
Discourses of refugees and asylum seekers in the UK press, 1996-2006. ESRC.

- October 2005 – March 2007.
- PI: Paul Baker; CIs: Tony McEnery, Ruth Wodak; RAs: Costas Gabrielatos (CL), Majid KhosraviNik (CDA), Michal Krzyzanowski (CDA).
- Corpus: 175,000 articles; 140 million words
- <http://ucrel.lancs.ac.uk/projects/rasim/>

Identifying spikes: Related issues

Trends vs. Outliers

Annual frequency development of *would* in *TIME*, 1923-2006
(from Millar, 2009: 201)

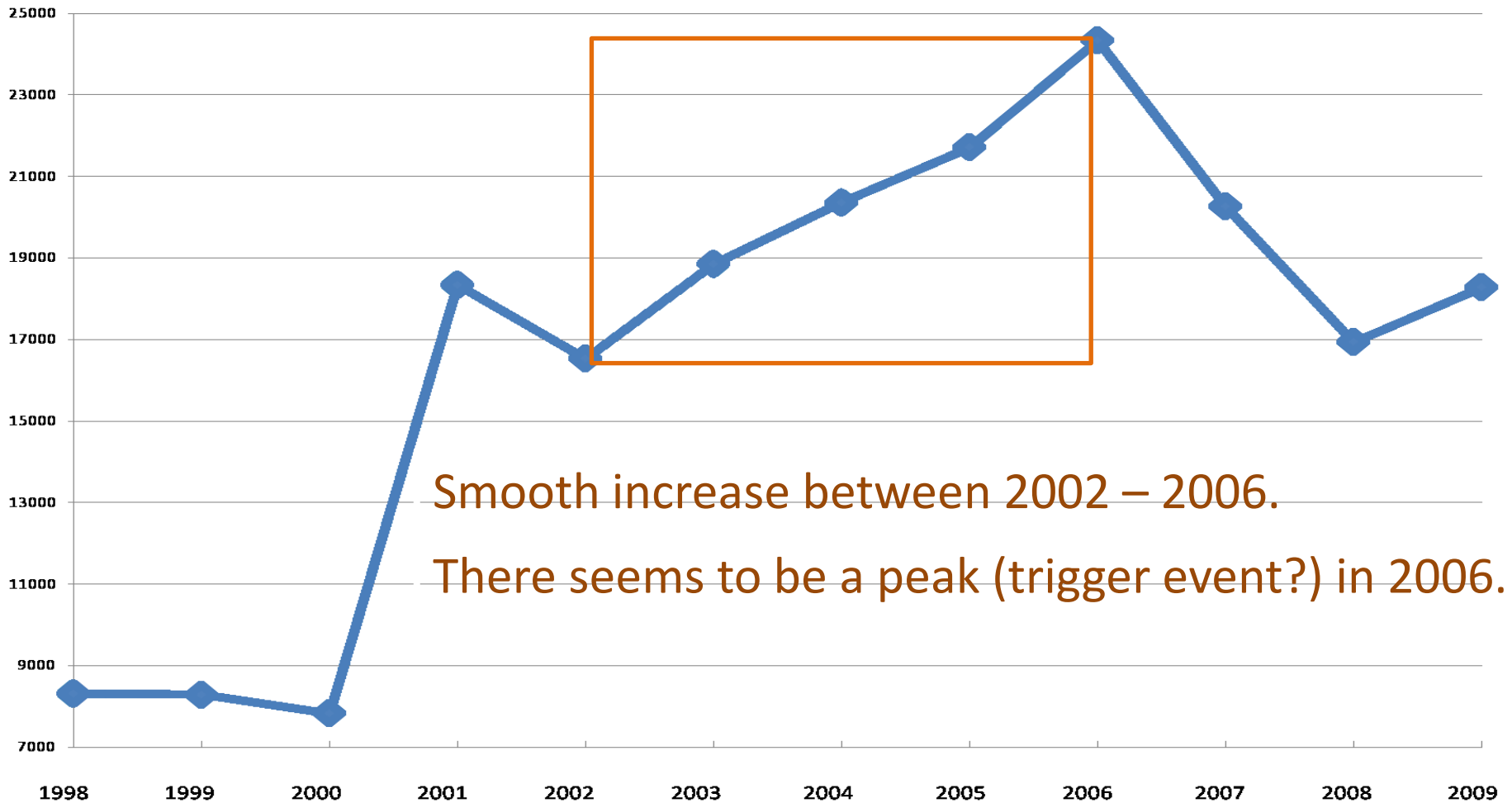


Time-span, sampling points, granularity

- *Time span*
 - No optimum span across the board: depends on aims of study
- *Sampling points*
 - Number of time-points at which frequencies have been measured
 - Related to *granularity* (Davies, 2010: 448) ...
 - ... but is not a useful indicator on its own.
- *Granularity*
 - The extent of detail that a given time span has been examined.
 - Metric needs to account of time-span *and* number of sampling points.
 - Metric needs to allow for granularity comparisons between diachronic corpus studies.

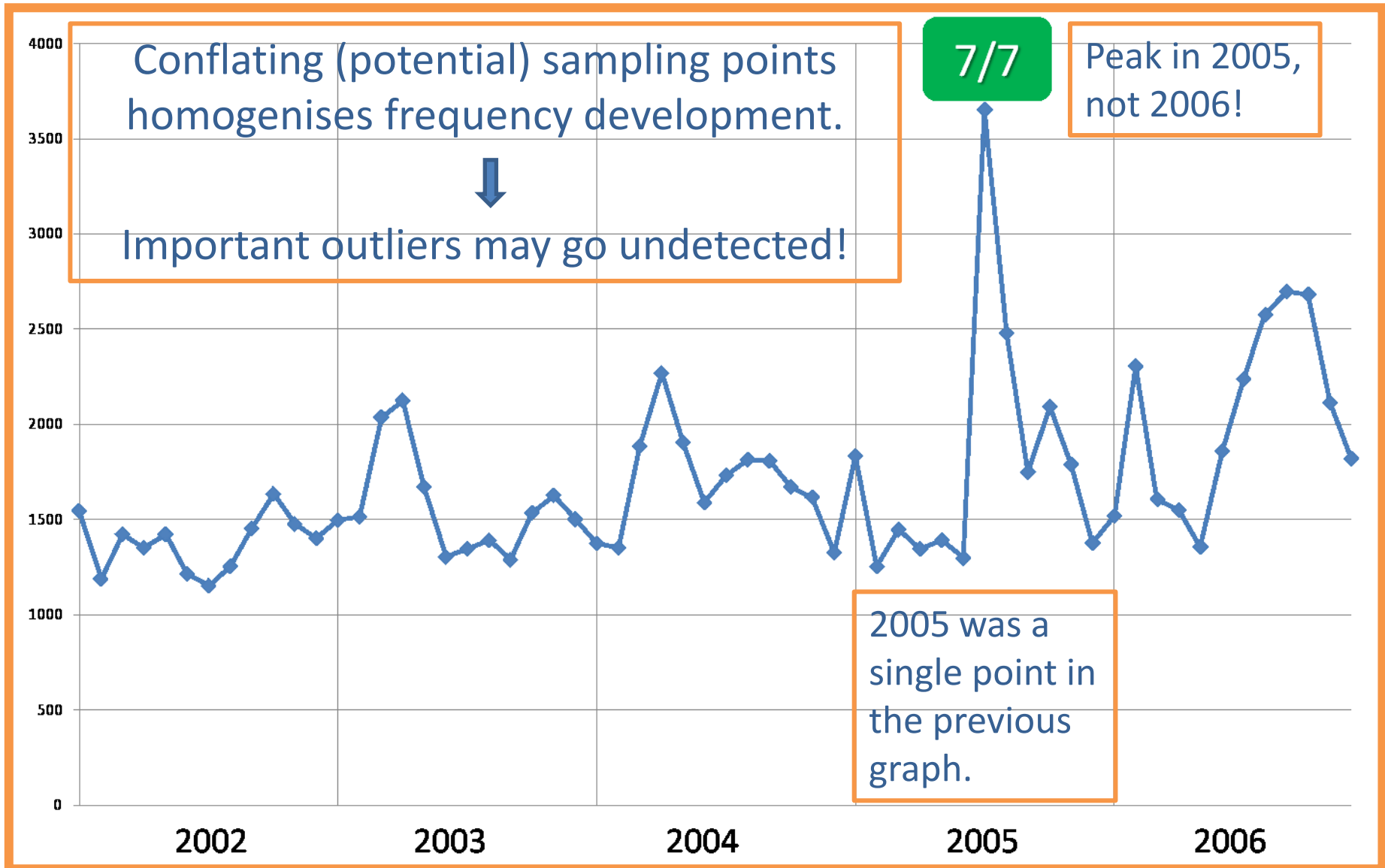
Sampling points and granularity

Islam corpus: article counts per year



Zooming in, adding granularity

2002-2006 article counts per month



Granularity

- Comparing the granularity of diachronic corpus studies.
 - Quantifying
 - Normalising

$$\frac{\text{Sampling points}}{\text{Time span (years)}}$$

Corpora	Sampling points	Time span	Granularity
Brown Family	4	80 (1928-2007)	0.05
<i>TIME</i>	84	84 (1923-2006)	1
Islam corpus	144	12 (1998-2009)	12

Context: CDA and CL

Context and CDA

CDA (particularly the Discourse-Historical approach) takes into account the following elements:

- The immediate co-text (e.g. collocations and resulting semantic prosodies).
- Intertextual and interdiscursive relations.
- Relevant contextual elements; e.g., “the occasion of the communicative event”.
- The broader sociopolitical and historical context.

(Reisigl & Wodak, 2001: 40-41)

Context and CL

One of the two main criticisms of CL (by CDA researchers):
CL does not take account of the relevant context

A non-linguistic quantitative corpus analysis (e.g. number of articles per period) reveals patterns which ...

- ... pinpoint periods/sources/texts that can be usefully examined in detail.

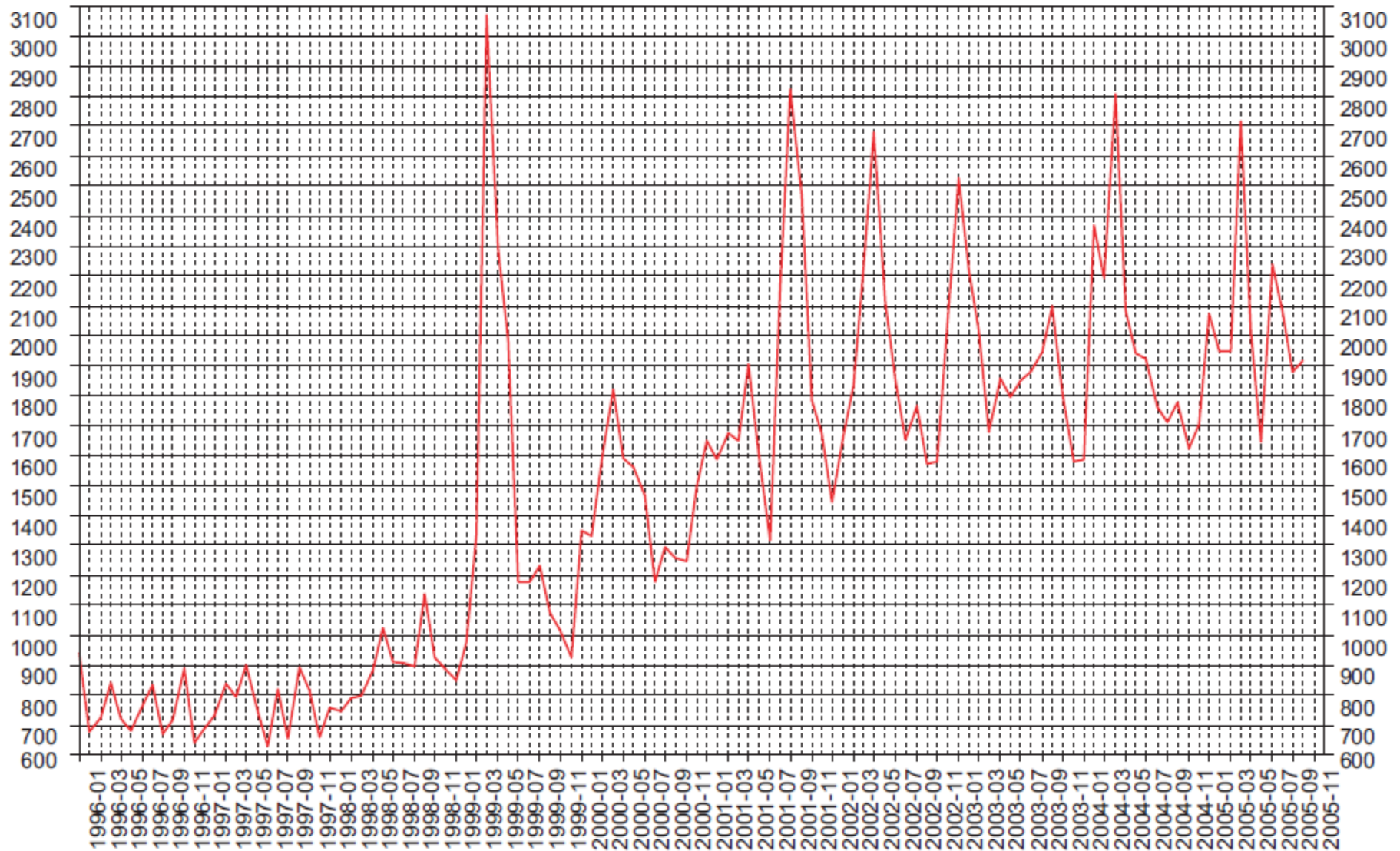
- ... uncover helpful contextual elements, which can inform the interpretation of results.

(Baker et al., 2008: 284-285)

CL researchers have no less access to sources of relevant contextual information than CDA researchers.

(Gabrielatos, 2009)

RASIM corpus: article counts per month



(Gabrielatos & Baker, 2008: 18)

RASIM: Spikes and candidate trigger events

Period	Events
March-May 1999	War in Kosovo Fighting between separatist guerrillas and paramilitary forces in East Timor
September-October 2001	Twin towers attack U.S. attack on Afghanistan Australian “boat people” incident
April-May 2002	Siege of the Church of the Nativity for 38 days. ¹⁵ War in Afghanistan East Timor independence
December 2002-February 2003	Chechen suicide truck-bomb attack Iraq disarmament crisis
March-April 2004	Second round of French presidential elections. The Asylum Bill in the United Kingdom. EU expansion-related immigration checks scandal. Madrid explosions Palestinian suicide bombers Violence in Kosovo Darfur ceasefire Assassination of Pim Fortyun.
March-May 2005	UK general elections Earthquake in Sumatra

Utility for CDA (RASIM project)

The identification of major spikes and the corresponding trigger events proved to be useful to the CDA strand of the RASIM project, because it ...

- “makes the data selection sensitive to the aims of deconstructing the representation of RASIM in the context of relevant socio-political developments” (KhosraviNik, 2009: 482).
- “helps the CDA strand to apply a preliminary restrictive factor in downsampling the texts” (KhosraviNik, 2010: 5-6).
- See also Baker et al. (2008: 284-285).

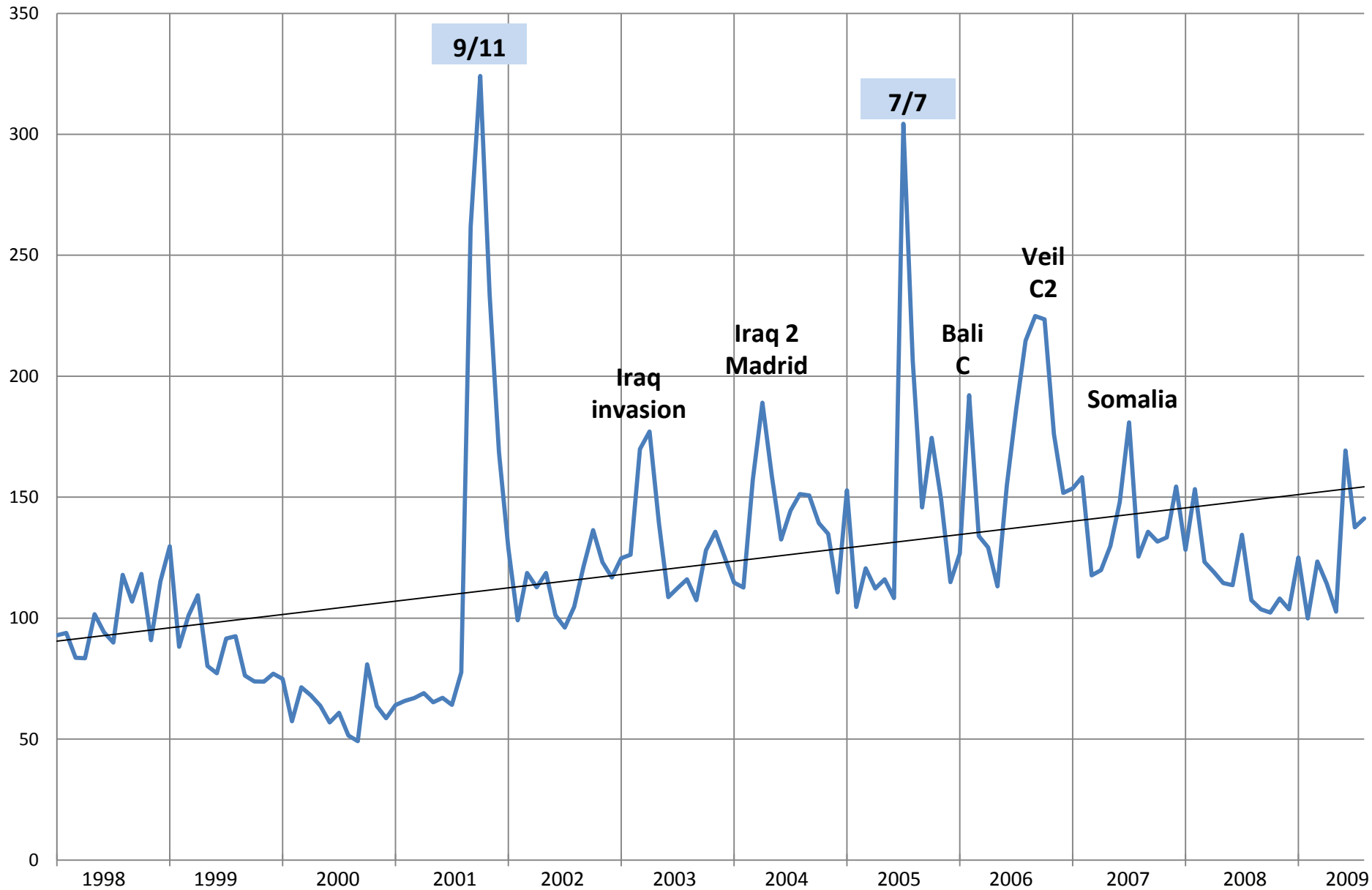
Limitations of the approach in the RASIM project

- The identification of spikes it was carried out in a limited and statistically naïve manner:
 - a) The frequency development of query terms was examined only in the whole corpus (i.e. all twelve corpus newspapers were taken collectively).
 - b) What was plotted was the average number of articles per month.
 - c) The spikes were established impressionistically.

(See also Gabrielatos & Baker, 2008: 17-20)

Whole corpus vs. Sub-corpora

Islam corpus: average number of articles per month



Spikes in Islam Corpus (as identified manually) and corresponding candidate trigger events

#	Spike date	Trigger Event	bu	ex	gd	in	ml	mr	ob	pp	st	su	tg	tm
1	1999, Jan.	<ul style="list-style-type: none">Iraq bombing in December 1998 (Operation Desert Fox).¹Christian-Muslim clashes in Indonesia.	✓											
2	2000, June-July	<ul style="list-style-type: none">Christian-Muslim clashes in Indonesia.Muslim rebels in the Philippines take hostages.	✓											
3	2000, Sept.-Oct.	<ul style="list-style-type: none">Muslim rebels in the Philippines take hostages.	✓											
4	2001, Sept-Oct.	<ul style="list-style-type: none">9/11.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	2002, March-April	<ul style="list-style-type: none">Hindus attacking Muslims in Western India.Discussion on the status of Muslims in the US.Reports on female genital mutilation in Muslim countries.Stoning of woman in Nigeria.							✓					
6	2003, March-April	<ul style="list-style-type: none">Invasion of Iraq.			✓	✓		✓		✓				✓
7	2004, March-April	<ul style="list-style-type: none">First anniversary of Iraq invasion.Madrid bombing.		✓						✓		✓		
8	2005, July-Aug.	<ul style="list-style-type: none">7/7.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
9	2005, October	<ul style="list-style-type: none">Bali bombings.Publication of Prophet Mohammed cartoons (C).	✓	✓	✓	✓	✓	✓	✓	✓	✓			
10	2006, Jan-Feb.	<ul style="list-style-type: none">Protests in Muslim countries in reaction to the publication of the Prophet Mohammed cartoons.					✓		✓			✓		
11	2006, Oct. (-Dec)	<ul style="list-style-type: none">Straw comments on Muslim women wearing the veil(Veil).²Prophet Mohammed cartoons – first anniversary (C2).	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
12	2007, Jan.-Feb.	<ul style="list-style-type: none">Somalia: Islamic rebels (Union of Islamic Courts) vs. government forces.									✓	✓		
13	2007, July	<ul style="list-style-type: none">No particular relevant event stands out.					✓					✓		
14	2007, Oct.-Dec.	<ul style="list-style-type: none">No particular relevant event stands out.								✓	✓			
15	2008, Feb. - March	<ul style="list-style-type: none">Archbishop’s comments on Sharia law.³							✓			✓		
16	2008, July	<ul style="list-style-type: none">No particular event stands out in frequency.										✓		
17	2009, March-April	<ul style="list-style-type: none">Obama visits Turkey.			✓							✓		
18	2009, June	<ul style="list-style-type: none">Iran elections.			✓							✓		✓
19	2009, August	<ul style="list-style-type: none">Flogging penalty for alcohol drinking in Malaysia.Clashes in Thailand and Somalia.							✓					

^[1] The main link to Islam is that the bombing was carried out just before Ramadan.

^[2] Jack Straw, member of parliament, leader of the House of Commons and ex-foreign secretary, said in a newspaper article that the veil is a "visible statement of separation and of difference" and that the practice of wearing it would make community relations in the UK harder.

^[3] The Archbishop of Canterbury, Dr. Rowan Williams, in a speech, supported the introduction of sharia law in the UK, because it could support social cohesion.

Is the general picture representative
of all corpus newspapers?

Yes

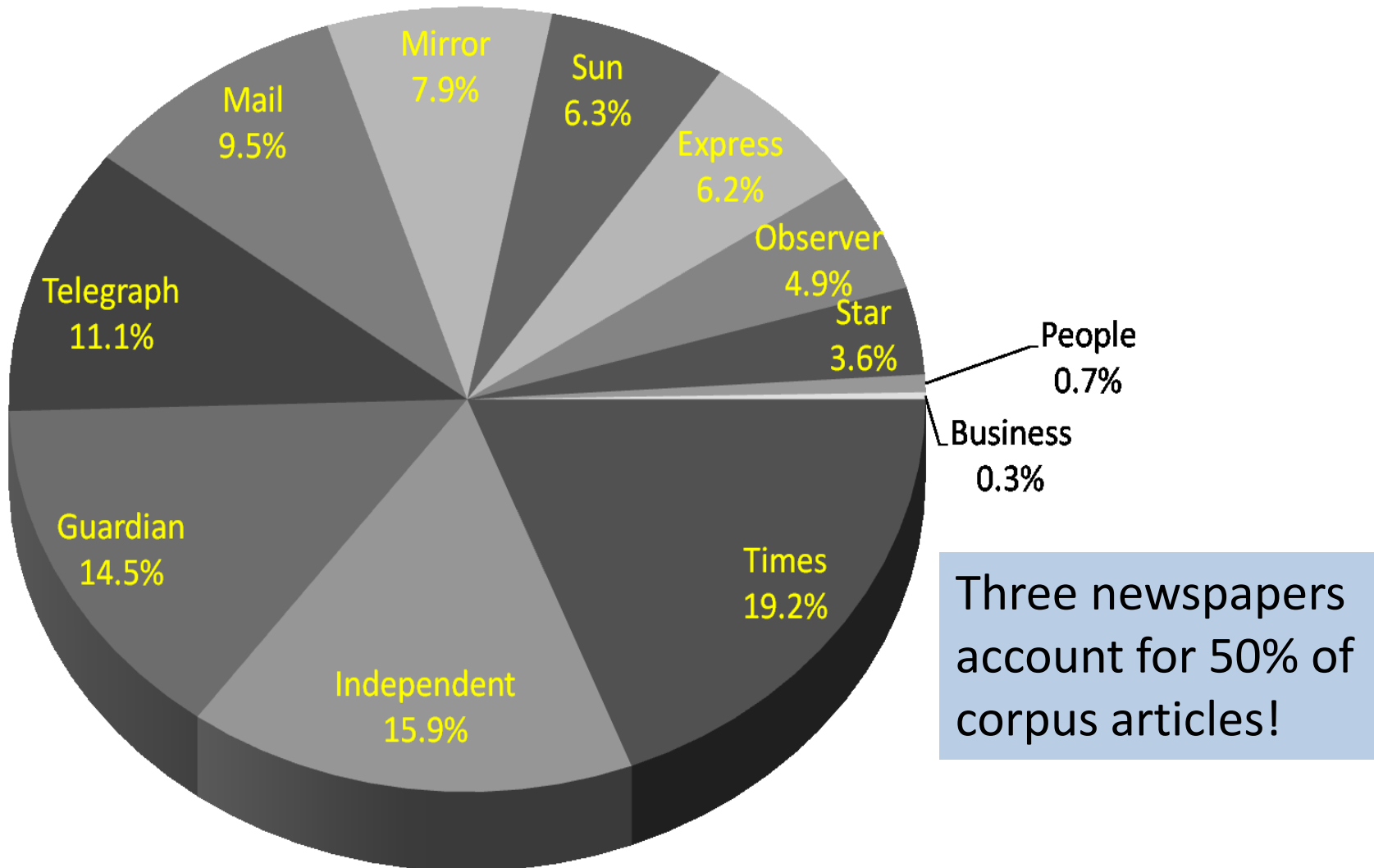
- Four spikes shared by at least two-thirds of the 12 newspapers:
 - 9/11, 7/7: 12
 - Veil + Cartoons 2: 11
 - Bali bombings + Cartoons: 9
- All newspapers but one show an upward trend.

No

- 19 spikes collectively – only 5 shared by more than half!
- Different relative importance of primary and/or secondary spikes.
- Five groups in terms of primary spikes:
 - 9/11 & 7/7
 - 9/11
 - 7/7 + other
 - Other + 9/11 & 7/7
 - Other

Corpus newspapers do not contribute equally

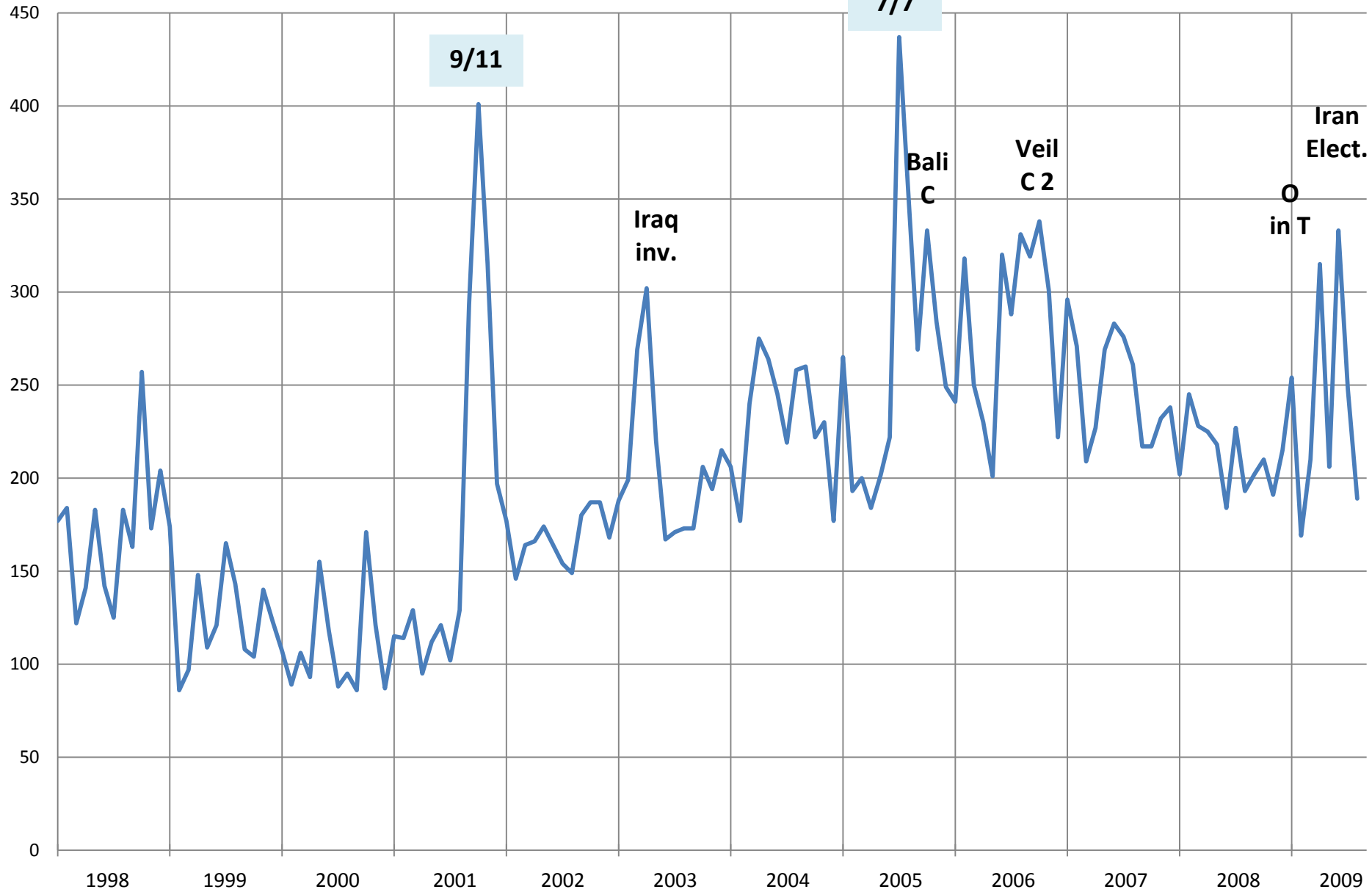
Islam corpus: Proportion of corpus articles from each newspaper



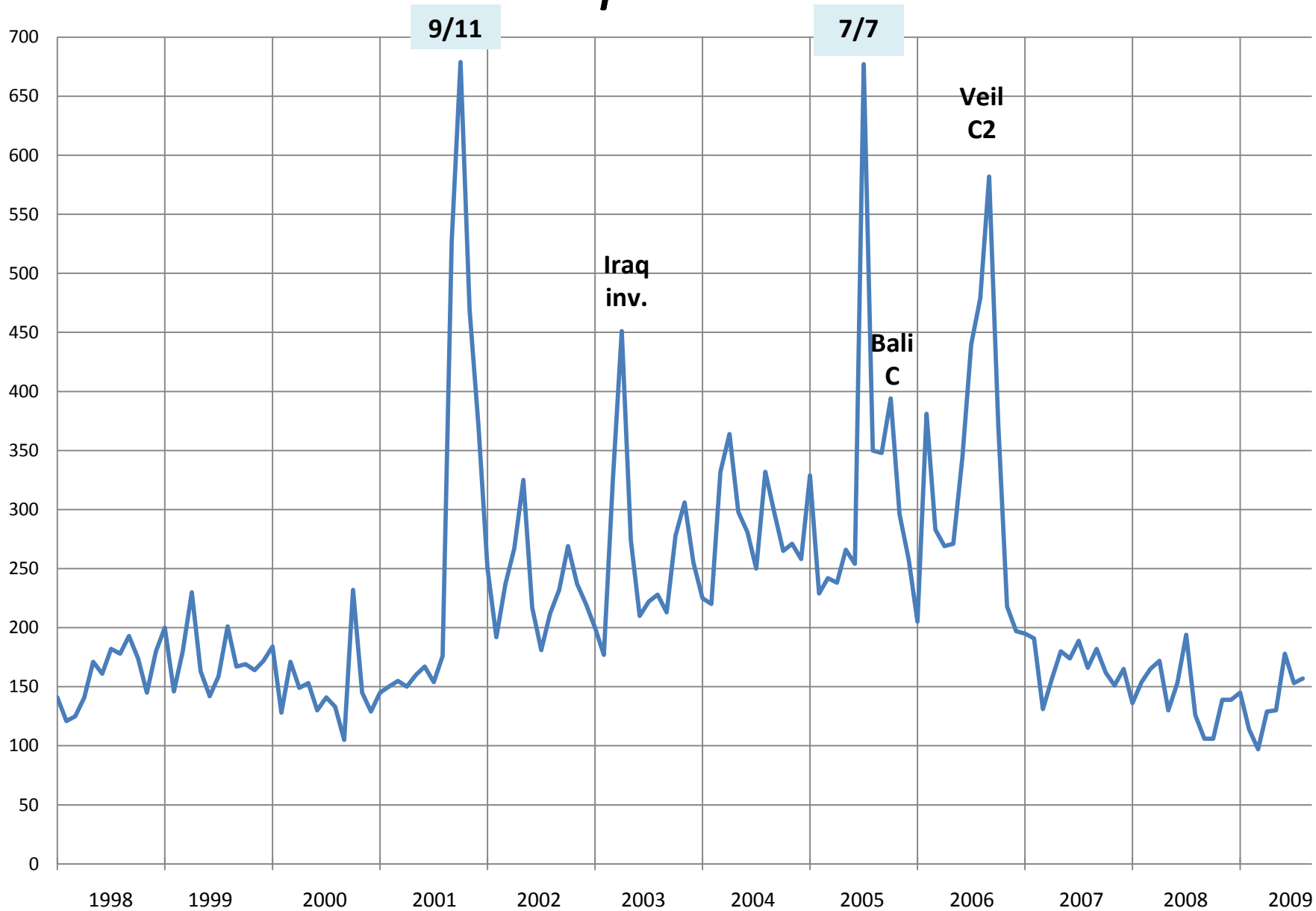
9/11 & 7/7

3 newspapers

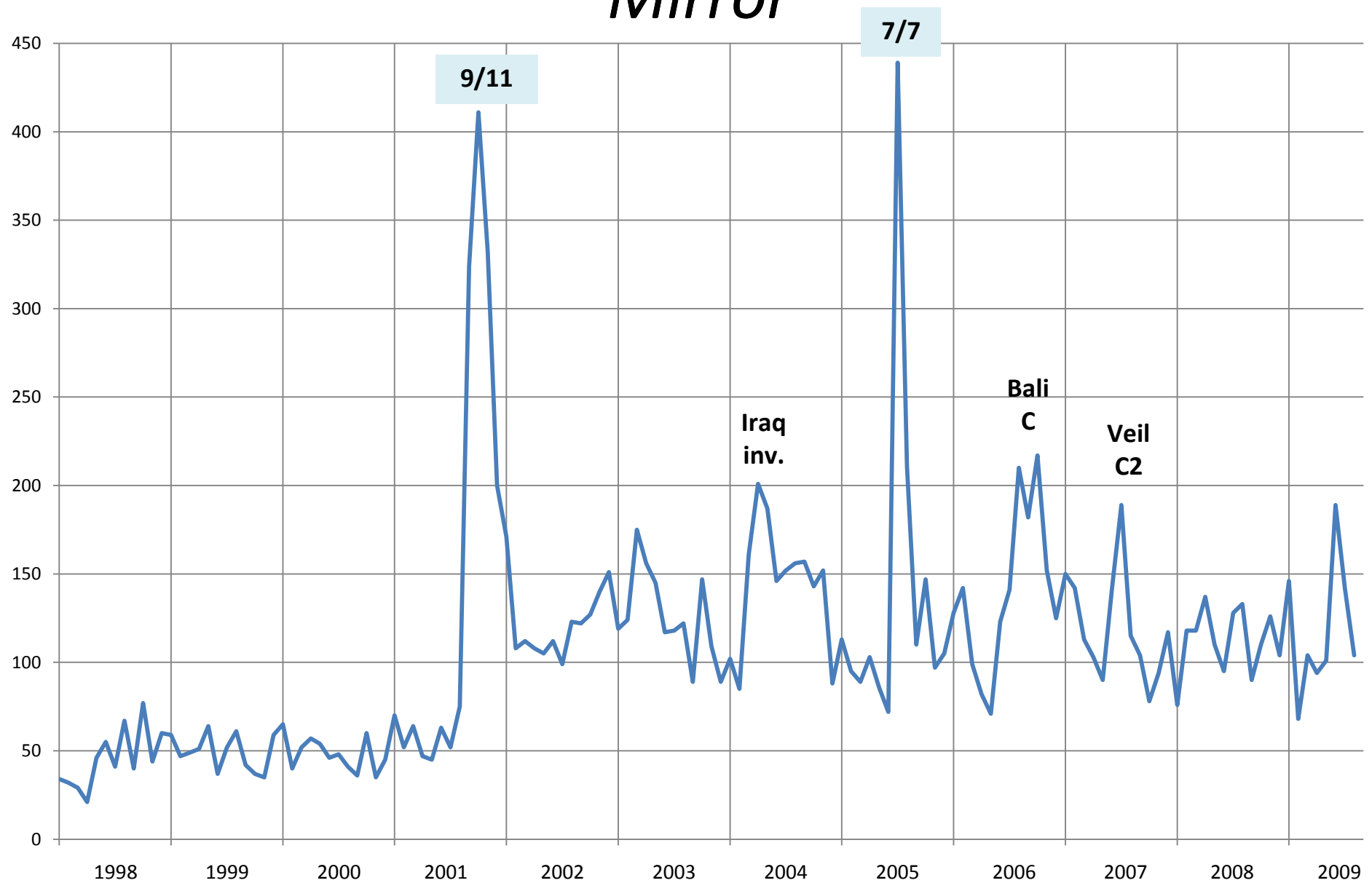
Guardian



Independent



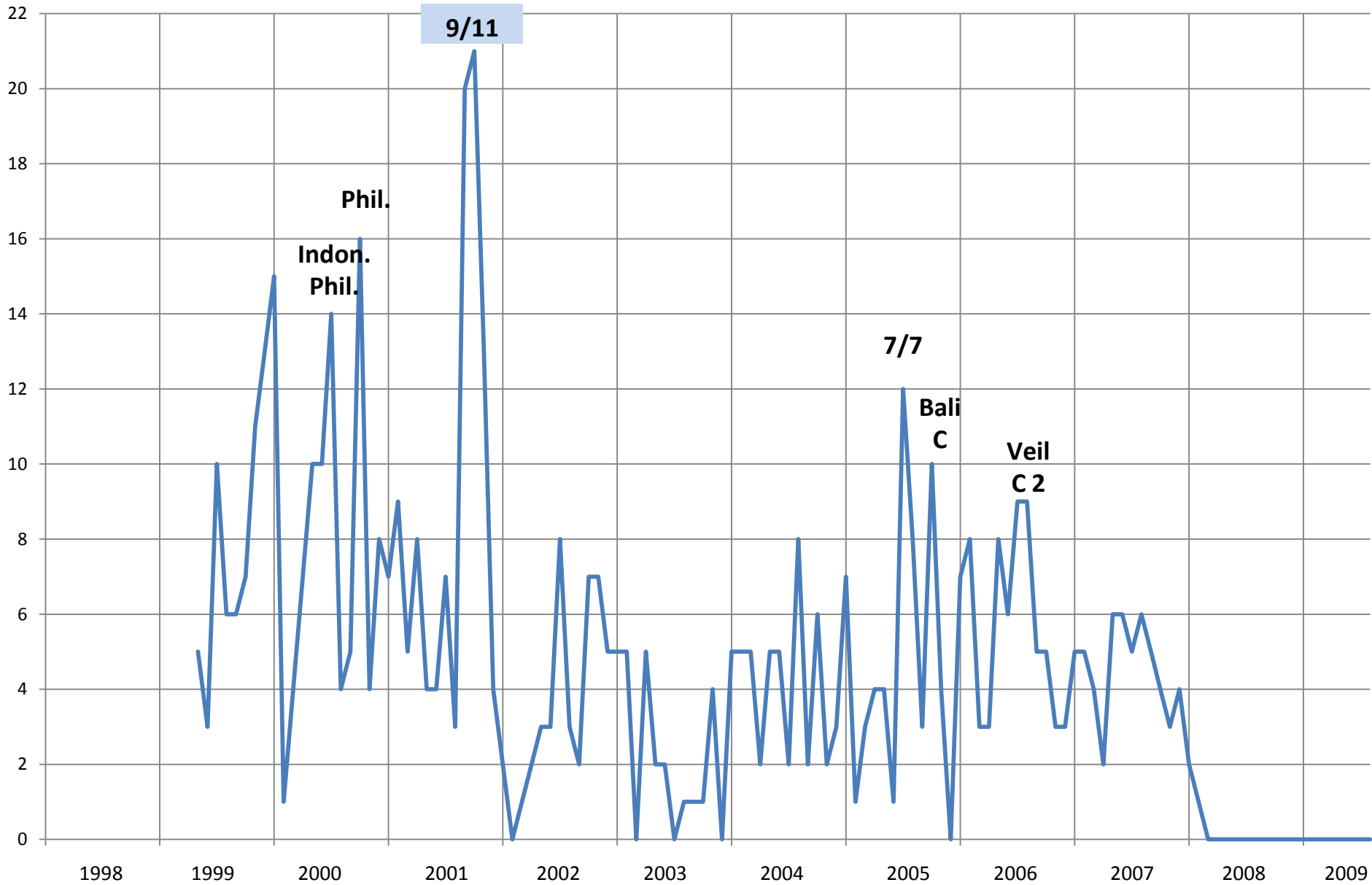
Mirror



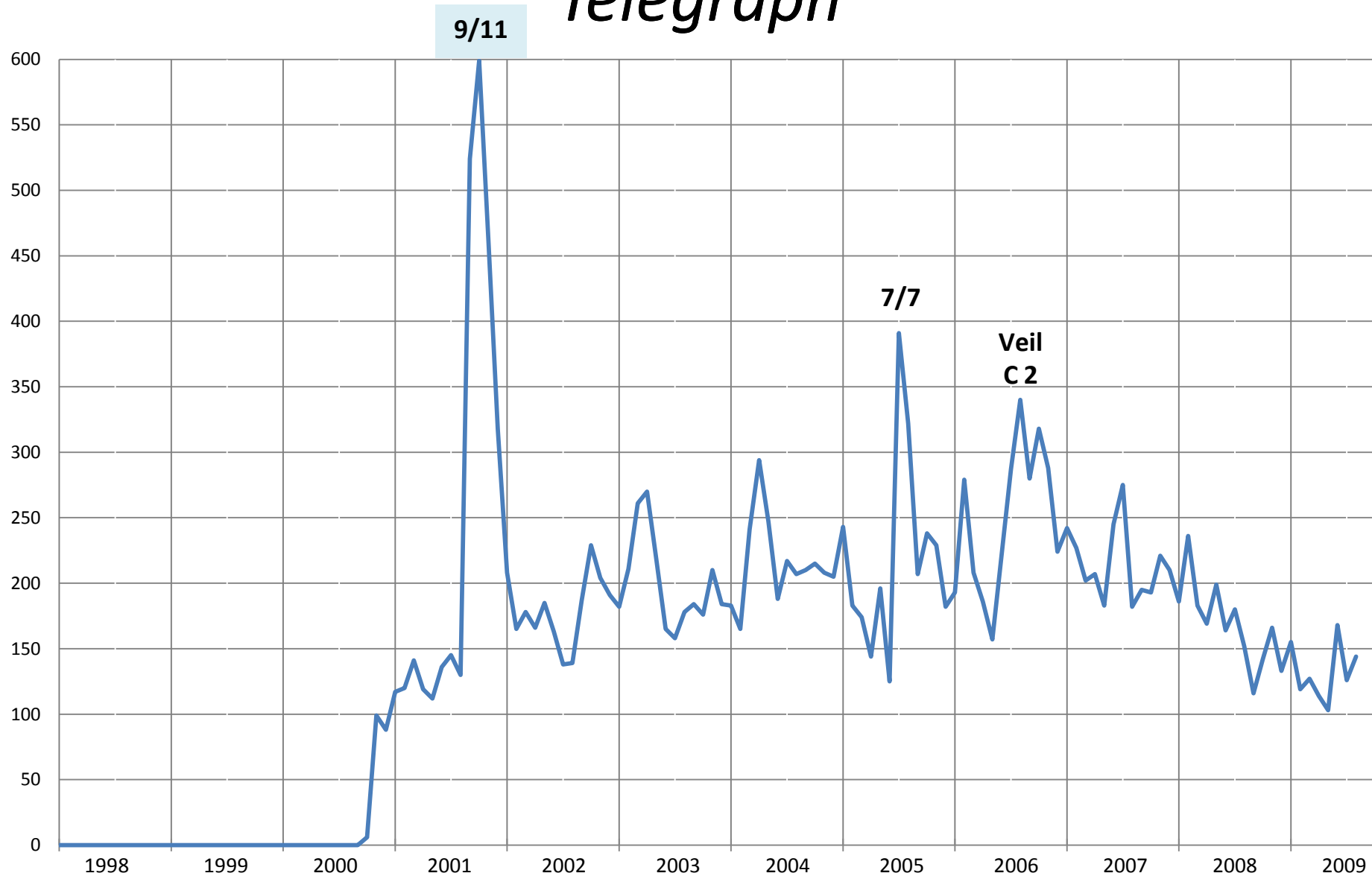
9/11

4 newspapers

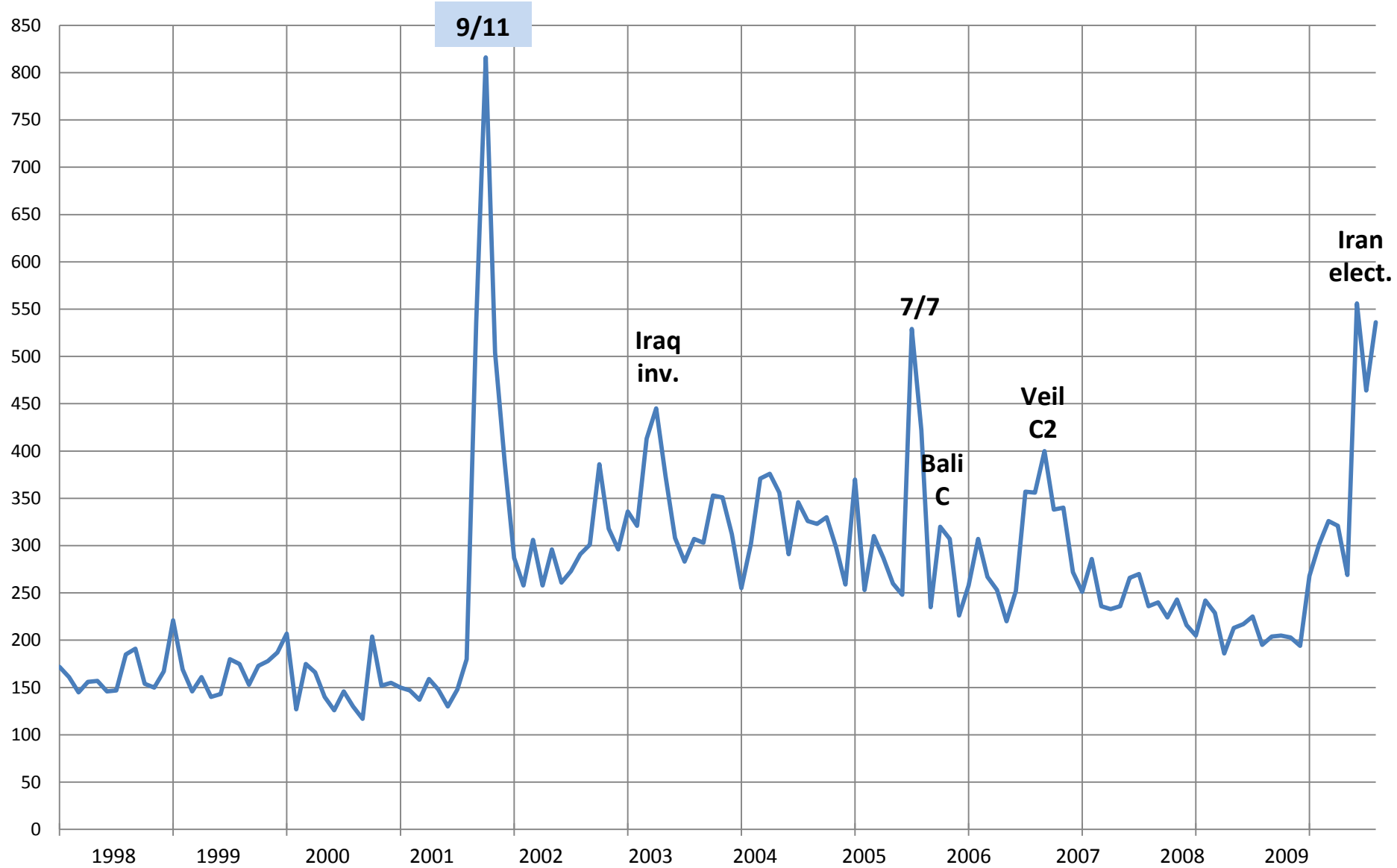
Business



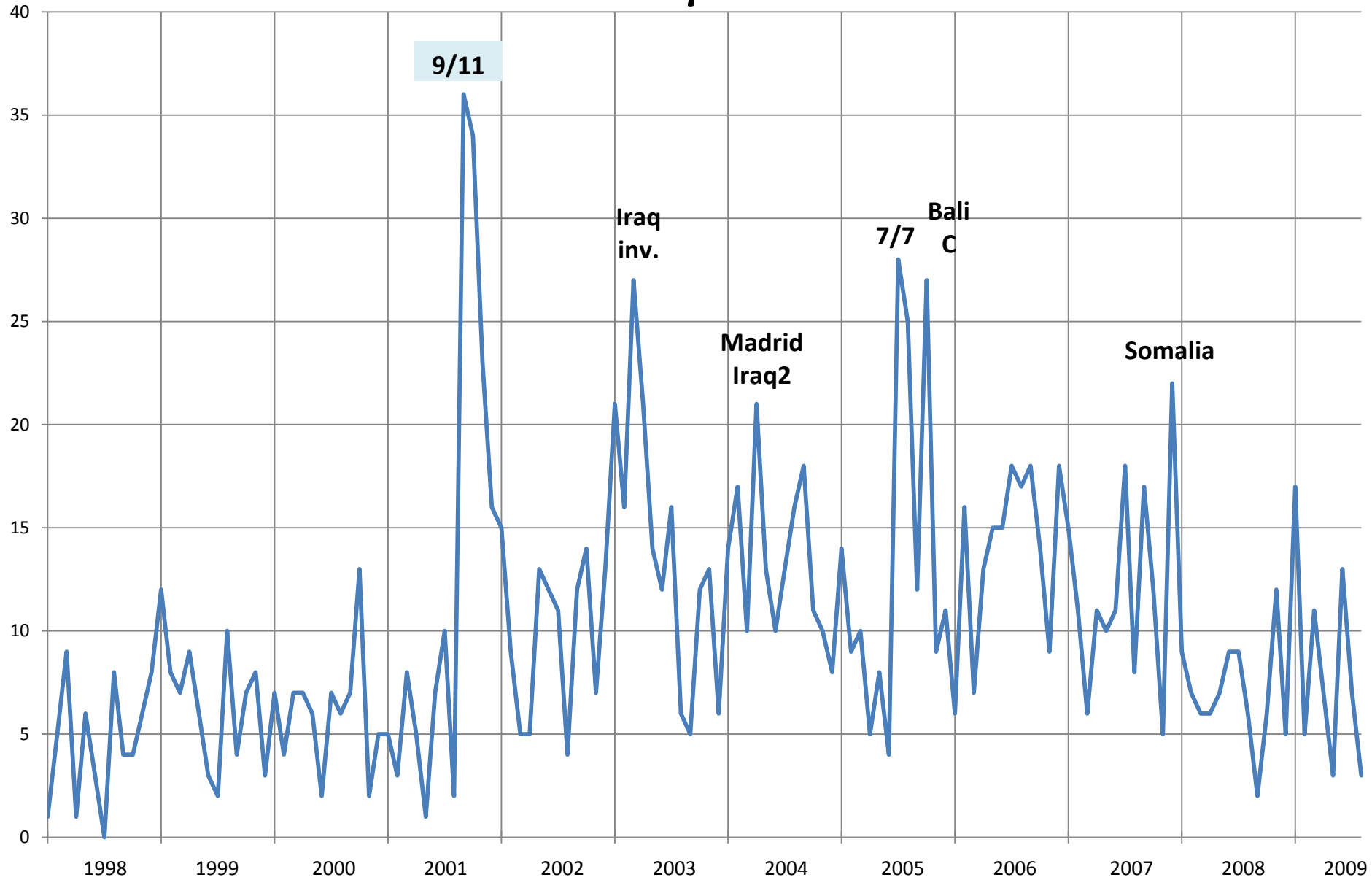
Telegraph



Times



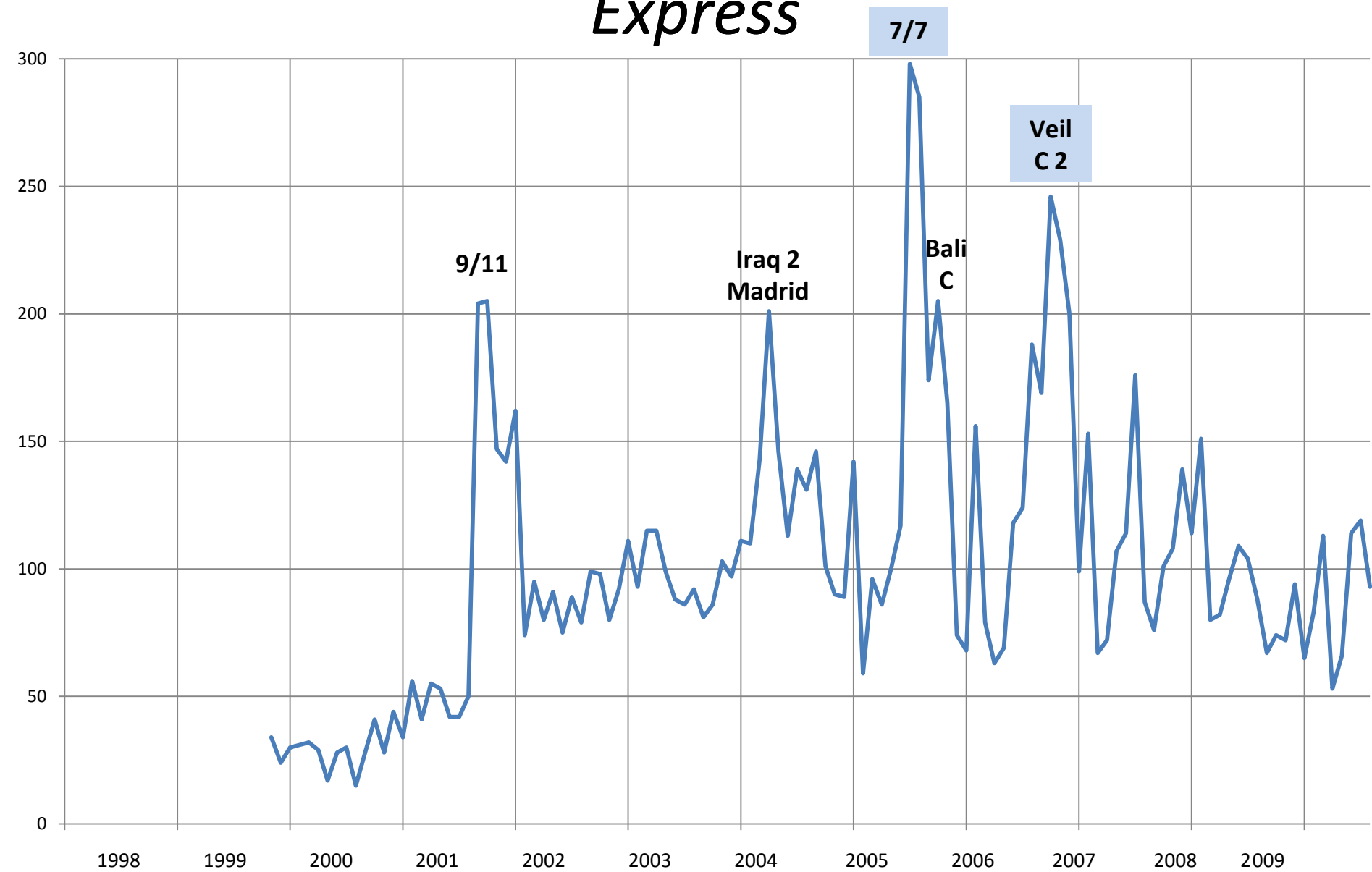
People



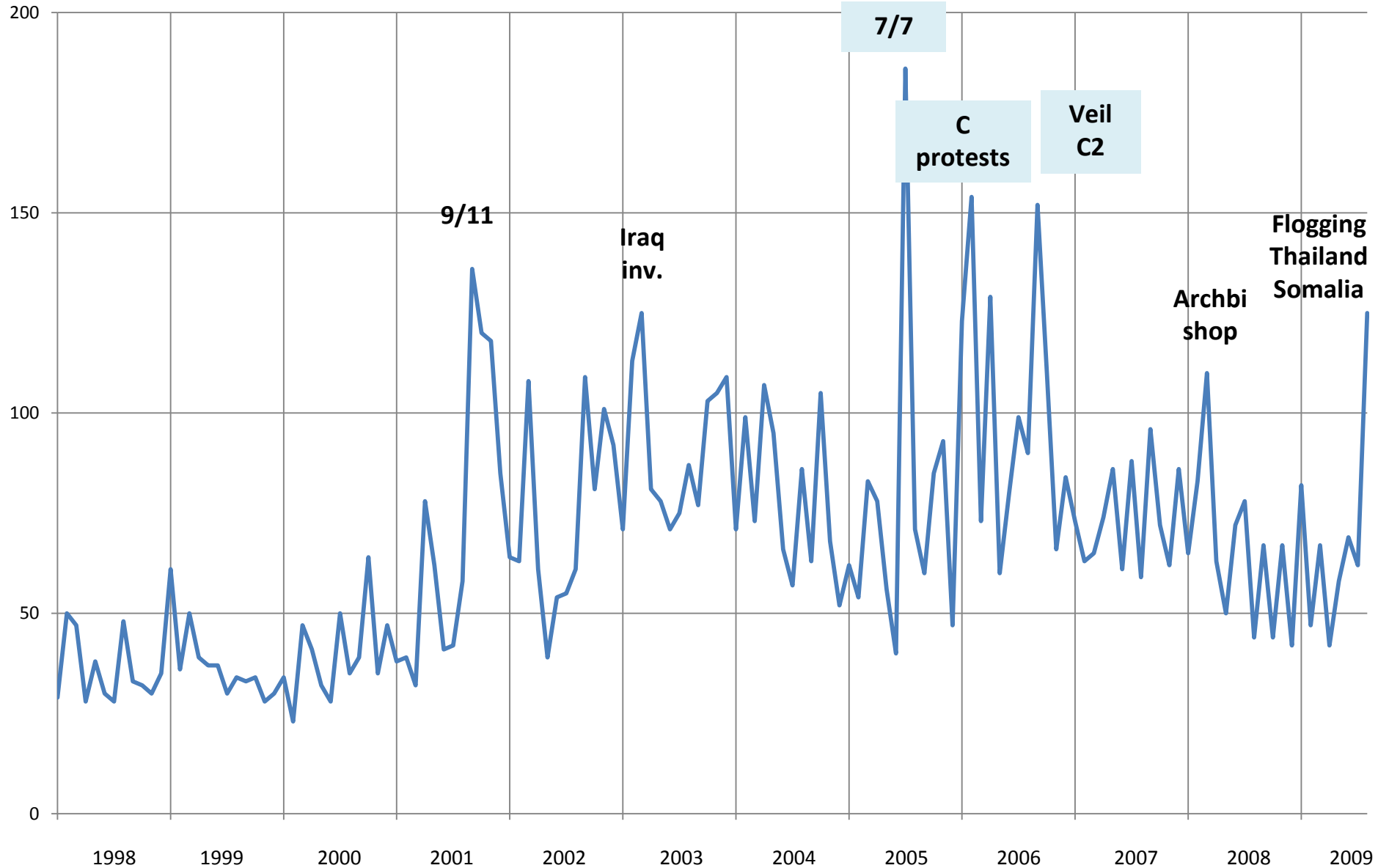
7/7
+
other

2 newspapers

Express



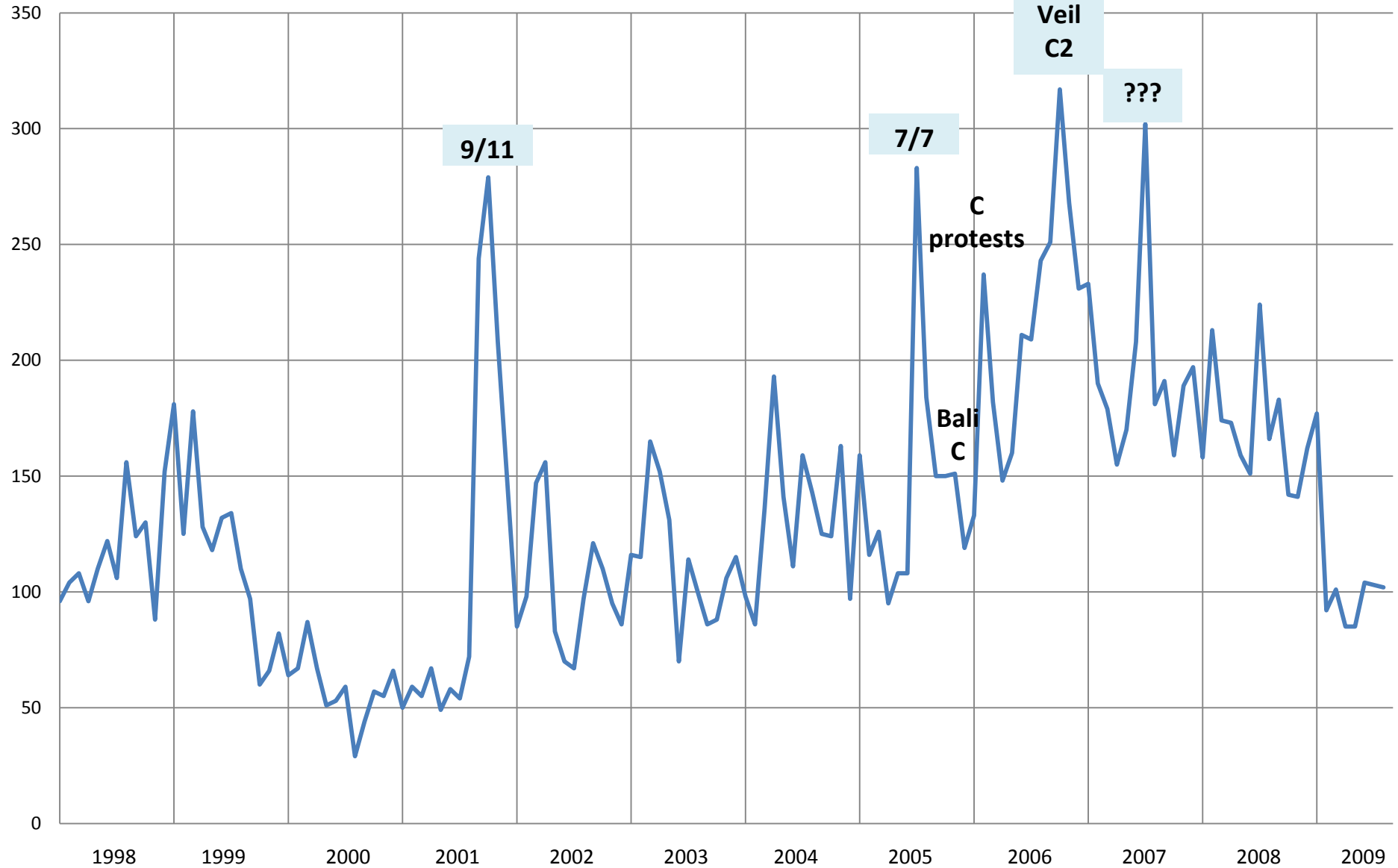
Observer



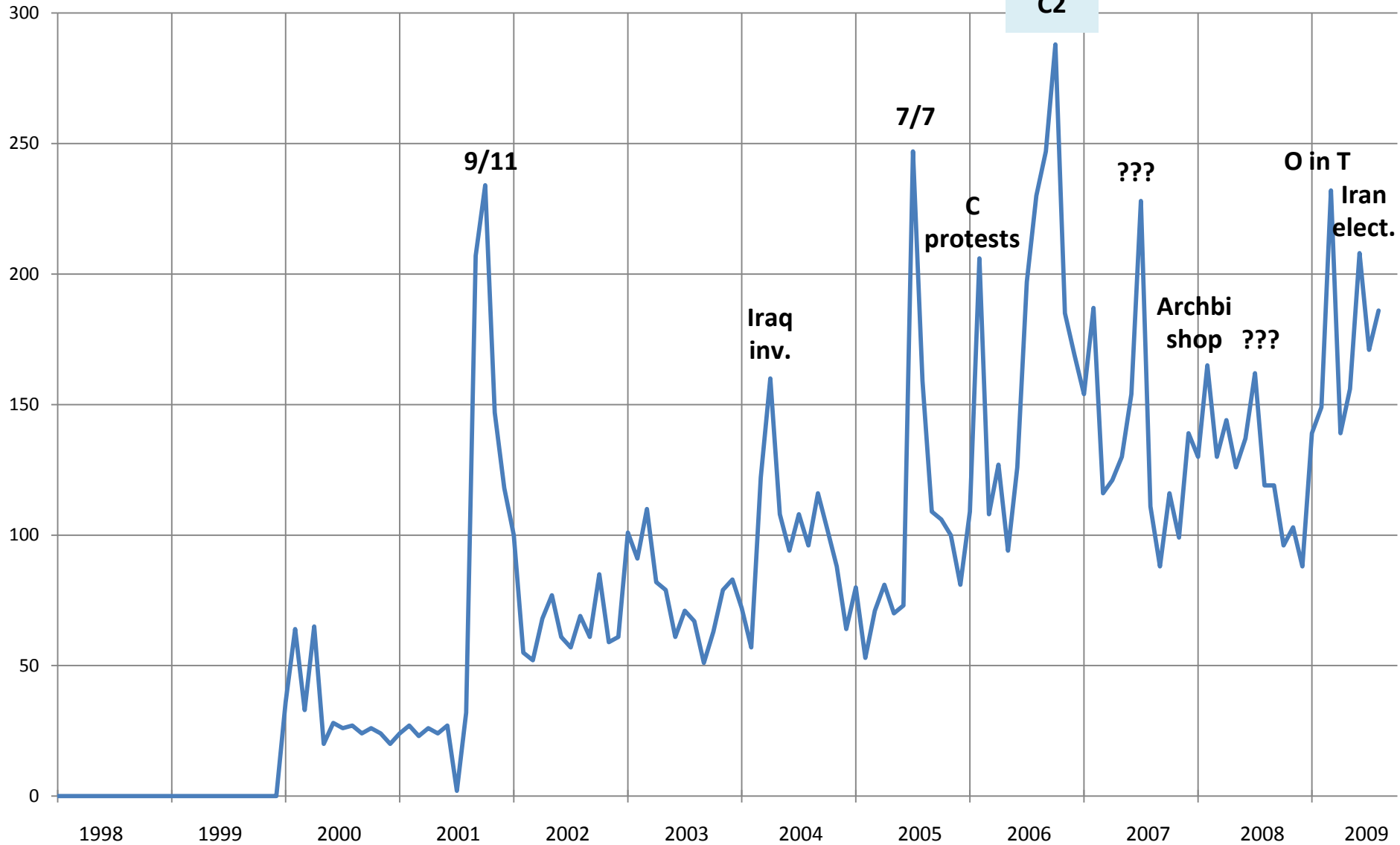
Other
+
9/11 & 7/7

2 newspapers

Mail



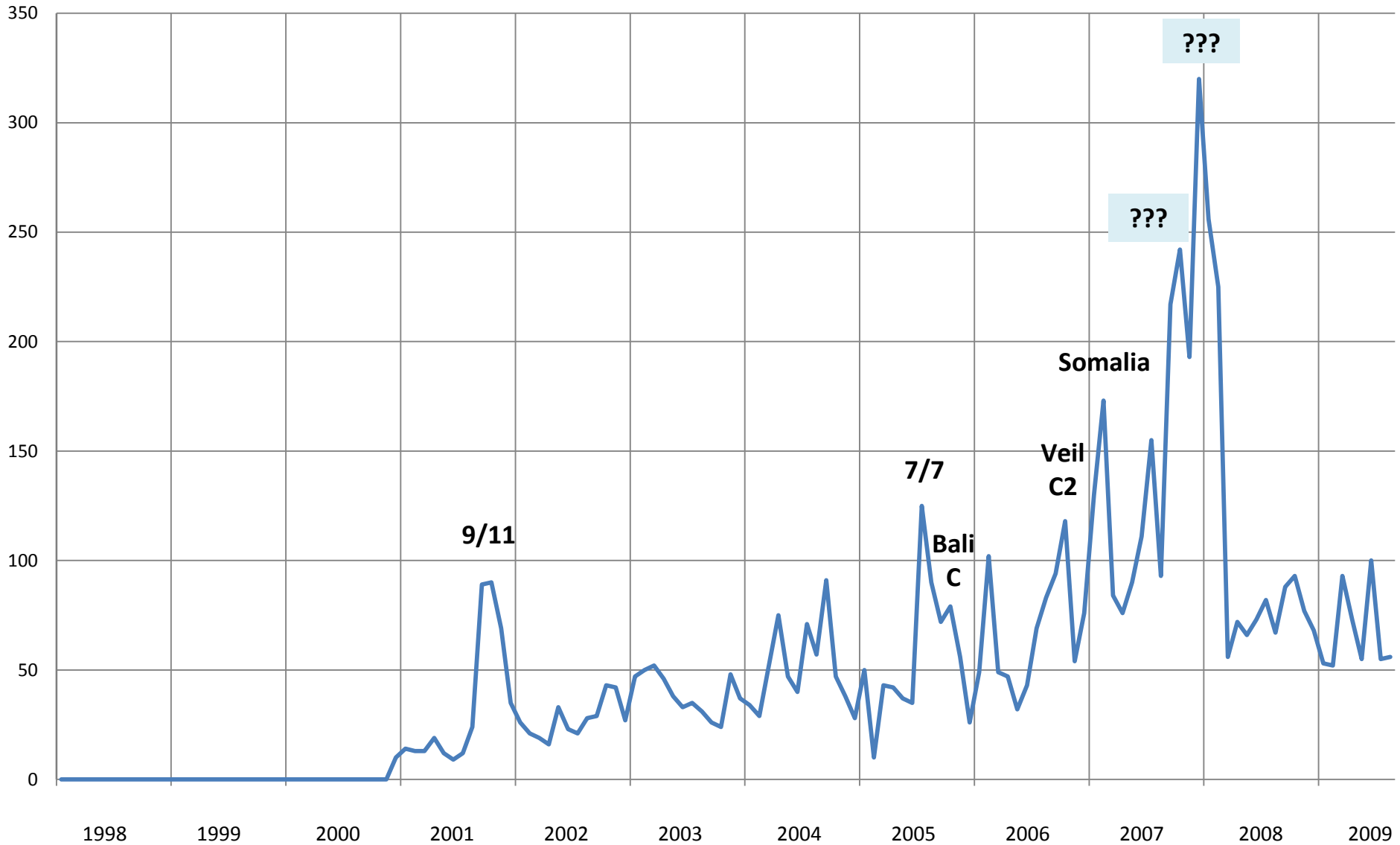
Sun



Other

1 newspaper

Star



A statistical approach to identifying outliers:

Wave, spike and trough (WST) method

Two current approaches in diachronic CL studies

- Focus on the statistical significance of trends over time.
 - Time points and the frequencies recorded at each point are treated as variables -- regression analyses establish the extent of correlation between the variables.
 - E.g., Baayen & Renouf (1996), Millar (2009).
- Focus on the statistical significance of frequency differences between two time points.
 - E.g. Leech & Smith (2006), Leech et al. (2009), Mair et al. (2003).
- Regression analyses and pairwise frequency differences seem to be adopted in studies of high and low granularity respectively.

WST method: Nature and calculations

- Combines the focus of current approaches.
- Successive pairwise frequency comparisons.
 - Similar to keyword comparison: The frequency of each sampling point is compared to that of the previous one .
 - Logarithm of frequency difference.
- Spikes starting from troughs are more significant than those starting from average-level or low-level spikes.
- Statistical significance established through a non-parametric regression model: $Y(t) = s(t) + Z(t)$
 - $s(t)$: function of time to be estimated.
 - $Z(t)$: a sequence of independent error terms.
 - The model was fitted using the mgcv package (Wood, 2006) within R.
- Two levels of statistical significance: 99%, 95%.

WST method: Findings (1)

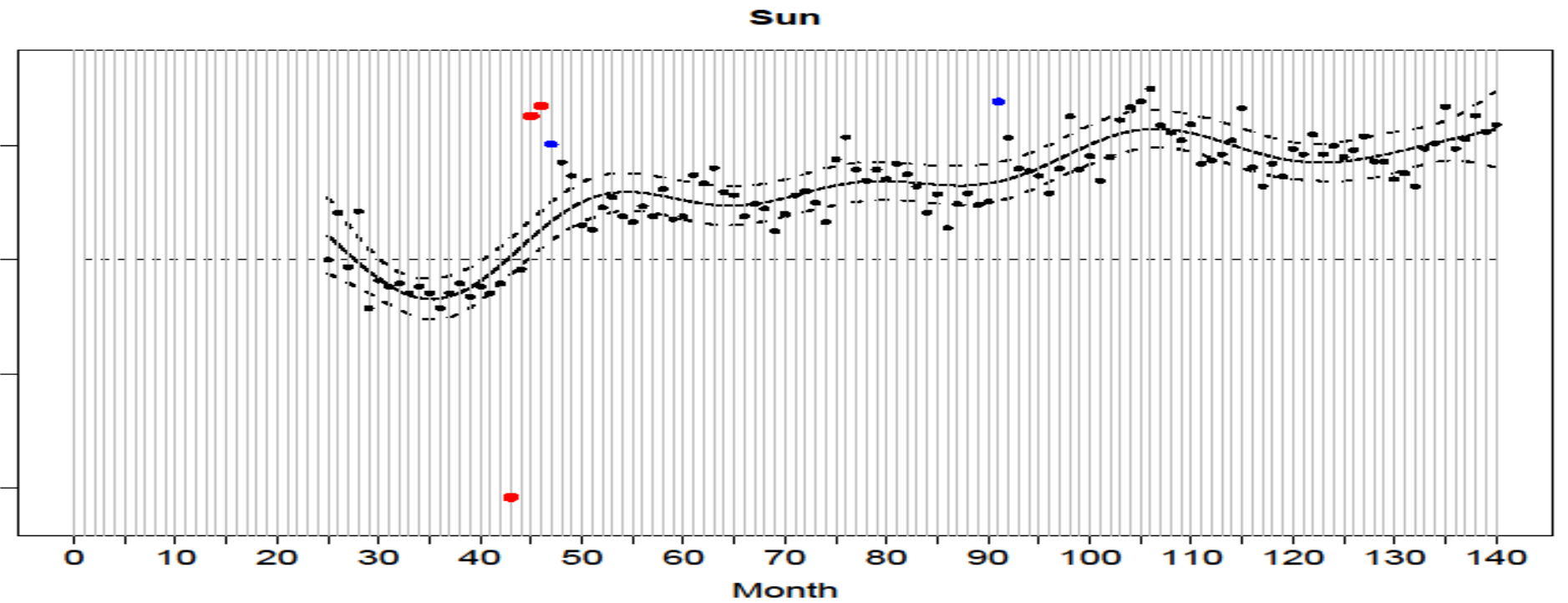
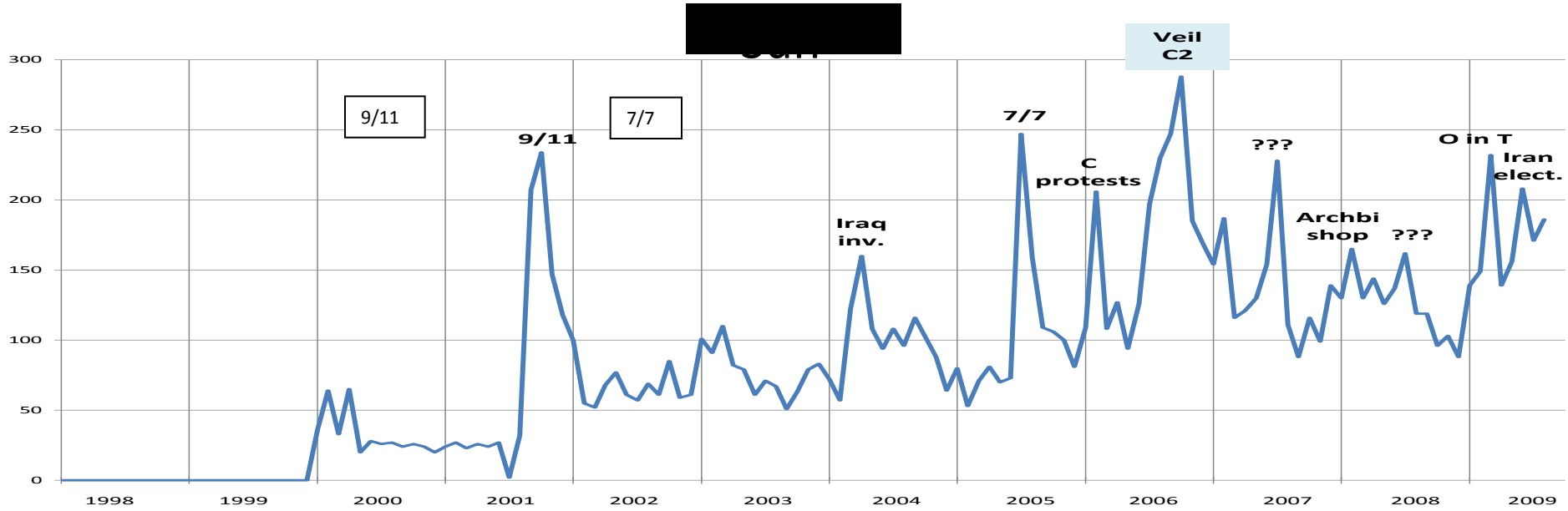
- Statistically significant spikes and individual trigger events were a subset of those identified manually.

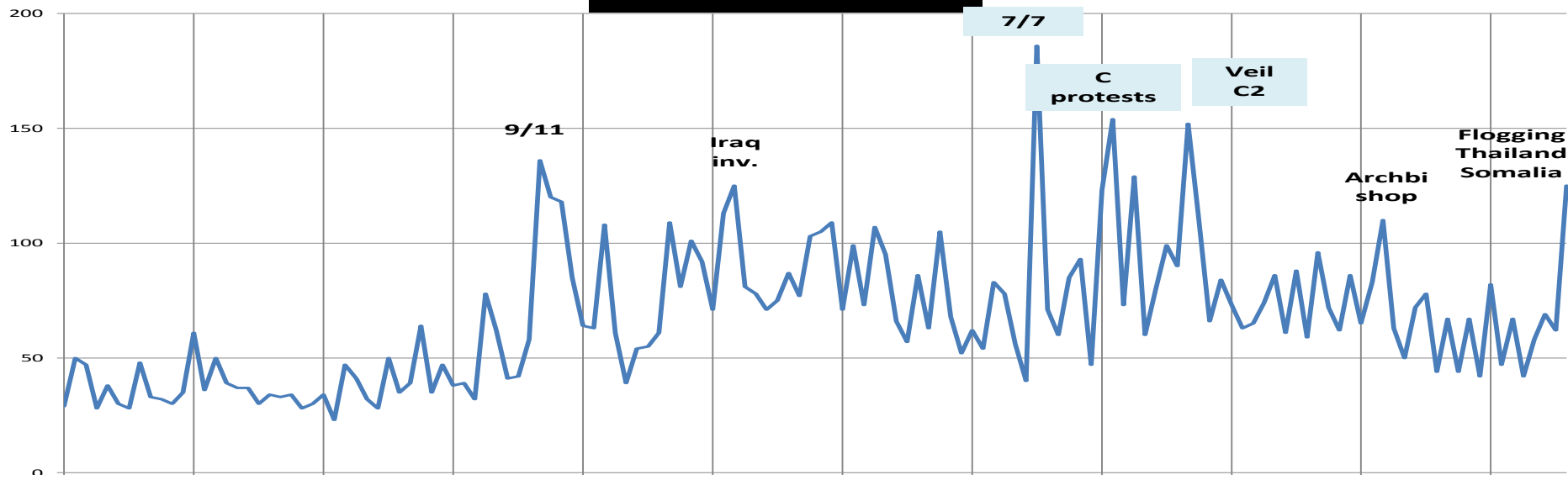
	Manual	WST
Spikes	74	29
Trigger events	19	9

- Only two spikes (trigger events) characterise UK national press as a whole: 9/11 and (to a lesser extent) 7/7.
- One significant spike unidentified by manual examination.
- One trigger event misidentified by the manual method (different newspaper).

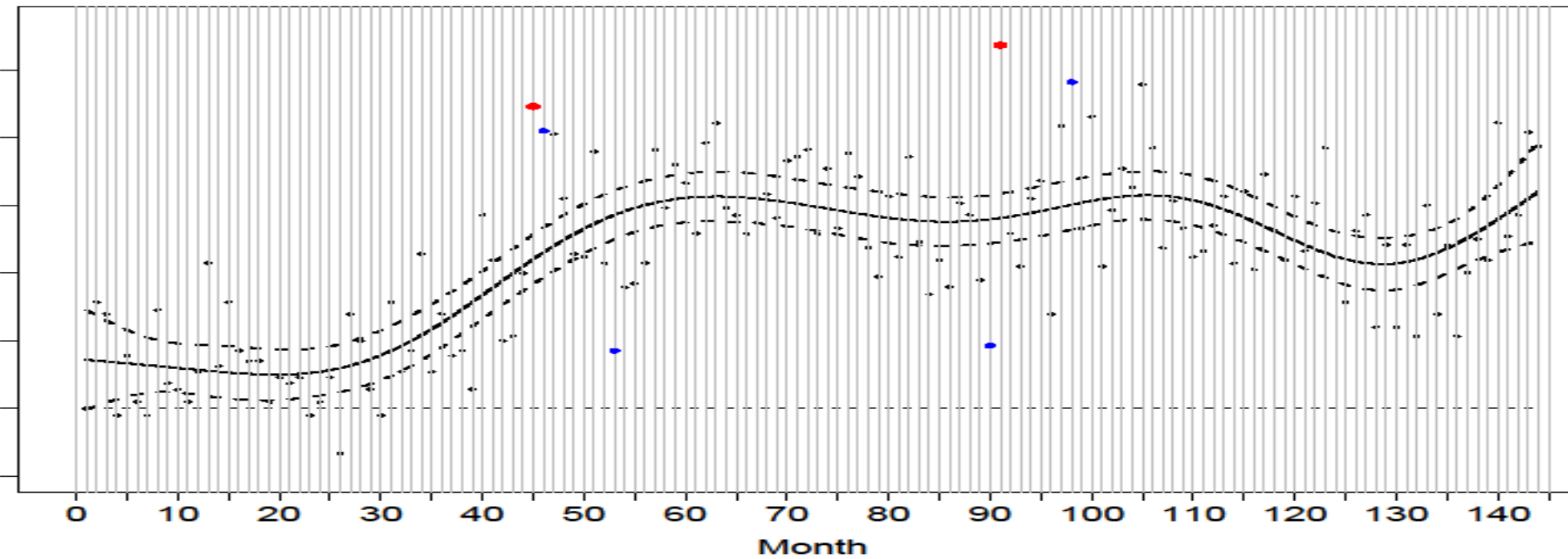
Comparison of approaches: spikes in Islam corpus

#	Spike date	<i>bu</i>	<i>ex</i>	<i>gd</i>	<i>in</i>	<i>ml</i>	<i>mr</i>	<i>ob</i>	<i>pp</i>	<i>st</i>	<i>su</i>	<i>tg</i>	<i>tm</i>
1	1998, Nov.-Dec.												
2	1999, January	✓											
3	2000, June-July	✓											
4	2000, Sept.-Oct.	✓											
5	2001, Sept-Oct.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	2002, March-April							✓					
7	2003, March-April			✓	✓		✓		✓				✓
8	2004, March-April		✓						✓		✓		
9	2005, July-Aug.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	2005, October	✓	✓	✓	✓	✓	✓	✓	✓	✓			
11	2006, Jan-Feb.					✓		✓			✓		
12	2006,Oct.(-Dec)	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
13	2007, Jan.-Feb.									✓	✓		
14	2007, July					✓					✓		
15	2007, Oct.-Dec.								✓	✓			
16	2008, Feb. - March							✓			✓		
17	2008, July										✓		
18	2009, March-April			✓							✓		
19	2009, June			✓							✓		✓
20	2009, August							✓					





Observer



WST method: findings (2)

Influential events raise the bar for subsequent candidate spikes.

- It will take an event of the same, if not greater, magnitude for a statistically significant spike to be established later, ...
- ... unless troughs in due course reduce the height of that notional bar.

Assessing the salience of contextual elements

Three complementary ways:

- The p -value of a given spike.
 - For example, spikes at the 99% level of confidence can be treated as more salient than those at the 95% level.
- The proportion of newspapers in which a particular event has registered a spike.
 - For example, in our study, 9/11 registered a spike in all newspapers.
- The number of spikes that a particular event registers, either overall in the corpus, or in particular newspapers.
 - E.g., events registering multiple spikes can be treated as more salient than those registering only one.
 - Clear example from our study are 9/11 and 7/7.

Summary

- Granularity is important for diachronic corpus studies.
- Corpus trends need to be compared to trends in sub-corpora.
- Identification of peaks points towards ...
 - time periods to be examined more closely.
 - time periods for CDA downsampling.
 - salient contextual elements to inform conclusions.
- The WST method establishes ...
 - statistically significant peaks and troughs,
 - diachronic frequency trends.
 - a much lower number of spikes and trigger events.

References

- Baayen, R.H. & Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language*, 72(1), 69-76.
- Baker, P., Gabrielatos C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273-305.
- Gabrielatos, C. (2009). Corpus-based methodology and critical discourse studies: Context, content, computation. Invited presentation. Siena English Language and Linguistics Seminars (SELLS), University of Siena, 9 November 2009. Abstract and slides available online: <http://eprints.lancs.ac.uk/28460/1/SELLS-Gabrielatos-Nov2009.pdf>
- Gabrielatos, C. & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. *Journal of English Linguistics*, 36(1), 5-38.
- KhosraviNik, M. (2009). The representation of refugees, asylum seekers and immigrants in British newspapers during the Balkan conflict (1999) and the British general election (2005). *Discourse and Society*, 20(4), 477-498.
- KhosraviNik, M. (2010) The representation of refugees, asylum seekers and immigrants in the British newspapers: A critical discourse analysis. *Journal of Language and Politics*, 8(3), 1-29.
- Leech, G. & Smith, N. (2006). Recent grammatical change in written English 1961-1992: Some preliminary findings of a comparison of American with British English. In A. Renouf & A. Kehoe (Eds.), *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, 186-204. Also available online: http://www.lancaster.ac.uk/fass/doc_library/linguistics/leechg/leech_and_smith_2006.pdf
- Leech, G., Hundt, M., Mair, C. & Smith, N. (2009). *Change in Contemporary English: A grammatical study*. Cambridge: Cambridge University Press.
- Mair, C., Hundt, M., Leech, G. & Smith, N. (2003). Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7(2), 245-264.
- Millar, N. (2009). Modal verbs in TIME: frequency changes 1923-2006. *International Journal of Corpus Linguistics*, 14(2), 191-220.
- Reisigl, M. & Wodak, R. (2001). *Discourse and Discrimination: Rhetorics of racism and antisemitism*. London: Routledge.
- Wood, S.N. (2006). *Generalized Additive Models: An introduction with R*. Boca Raton: Chapman and Hall/CRC Press.